Credibility Scoring for Retrieval-Augmented Answers: Algorithm, Evidence, and Roadmap

Technical Report — Algorithmic Approach and Scientific Research Supporting the Credibility Score

Abstract

We specify a modular credibility-scoring system for a retrieval-augmented QA (RAG) pipeline. The score combines four signal families—source, content, evidence/entailment, and propagation/behavior—into a calibrated probability that a response (and its cited sources) are reliable. The design is grounded in established research on source trust estimation, claim verification, and misinformation detection, and informed by industry frameworks (e.g., E-E-A-T, NewsGuard) and community-moderation algorithms (e.g., bridging-based ranking). We present the rationale, complexity/scalability analysis, and an experiment plan (FEVER, LIAR, MultiFC) to validate performance and guide future iterations.

1. Algorithmic Approach & Rationale

1.1 System Overview

Given a user query, the RAG pipeline retrieves passages, generates an answer, and cites evidence. In parallel, the Credibility Scorer computes: (1) a Source Prior (S), (2) Content Signals (C), (3) Evidence/Entailment metrics (E), and optionally (4) Propagation/Behavior signals (P). These are fused via a calibrated model to output a probability $\hat{p} \in [0, 1]$ that the answer is credible. Thresholds map $\hat{p}$ to tiered labels (High, Medium, Low) with uncertainty qualifiers.

1.2 Feature Families

- A. Source (S) — domain prior (historical factual accuracy, corrections policy, ownership transparency, bylines), author expertise, link-graph hygiene; aggregated to $S \in [0, 1]$.
- B. Content (C) — stylometry/linguistic cues (clickbait markers, hedging), structure (citation density, quotes:claims ratio, entity density), claim extraction and temporal specificity.
- C. Evidence/Entailment (E) — per-claim retrieval against trusted corpora; NLI to classify entailed/contradicted/insufficient; aggregate entailment rate, contradiction rate, and retrieval confidence.
- D. Propagation/Behavior (P) — if social signals are available: early spread velocity, bot-likeness of amplifiers, cross-community diffusion asymmetry.

1.3 Scoring & Thresholds

Fuse signals via regularized logistic regression or gradient-boosted trees over [S, C, E, P], followed by probability calibration (Platt scaling or isotonic). Choose decision thresholds (t_high, t_low) on validation to satisfy product constraints such as precision@High ≥ 0.9. Quantify uncertainty via bootstrap over claims and sources; if confidence intervals span thresholds or evidence is insufficient, set state to “Insufficient Evidence”.

1.4 Complexity & Scalability

For m claims and k evidence items per claim over an ANN index of N passages, retrieval cost is ~O(m·k·log N). NLI inference dominates and is mitigated by batching, caching, and sharding. Source-prior computation is periodic offline; online lookups are O(1). Scale via pre-embedding corpora, FAISS/HNSW indexes, and async batching.

2. Literature & Industry Review

- Knowledge-Based Trust (KBT): estimate domain trust by correctness of extracted facts (informs Source Prior).
- Claim-verification datasets: FEVER (retrieval + entailment), LIAR (short political claims), MultiFC (multi-domain fact-checking).
- Misinformation detection surveys: cover stylistic, knowledge-based, and propagation features across classical ML and deep models.
- Propagation credibility: early Twitter studies highlighting user/content/propagation signals.
- Industry frameworks: E-E-A-T and NewsGuard provide interpretable rubrics; community-notes style bridging-based ranking mitigates factional capture.

Gaps addressed: over-reliance on style (we emphasize evidence + priors); black-box scoring (we expose calibrated probabilities & CIs); weak out-of-domain generalization (dataset mixing and domain priors).

3. Methodology Justification & Trade-offs

ML-based fusion — high accuracy and flexibility; requires calibration and drift monitoring.

Rule-based guardrails — transparent and low-latency; limited recall and brittle to edge cases.

Chosen hybrid — ML for scoring; hard rules for safety; optional human-in-the-loop with bridging-style aggregation.

4. Experimental Validation

4.1 Datasets & Splits

Use FEVER (train/dev/test) to calibrate entailment; LIAR for short political claims; MultiFC for robustness across domains.

4.2 Tasks & Metrics

- Binary credibility: AUROC, AUPRC, F1 at operating threshold.
- Calibration: Brier score, Expected Calibration Error (ECE), reliability diagrams.
- Claim-level: FEVER label accuracy, evidence precision/recall.
- Ablations: remove E, S, C, P modules to quantify contribution; cross-domain transfer tests.

4.3 Protocol

- Train fusion on FEVER(train) + LIAR(train).
- Calibrate on FEVER(dev) using isotonic regression.
- Tune thresholds on MultiFC(dev) for desired precision/recall trade-off.
- Evaluate on MultiFC(test) and LIAR(test); report metrics and ablations.
- Stress-test: remove top-prior sources and observe score drift.
- Human spot-check: 100 samples using an E-E-A-T-style checklist.

5. API & Implementation Details

5.1 Inference-time API (JSON over HTTP)

Request (JSON):

```
{
  "query": "string",
  "answer": "string",
  "citations": [{"url": "https://example.com", "title": "string", "author": "string", "published_at": "2025-09-01T
00:00:00Z"}],
  "enable_propagation_features": false,
  "return_explanations": true
}
```

Response (JSON):

```
{
  "score": 0.83,
  "label": "high",
  "confidence_interval": [0.77, 0.88],
  "explanations": {
    "top_features": [{"name":"evidence.entailment_rate","contribution":0.27}],
    "per_claim": [{"claim":"...","retrieval@k":20,"entailed":true,"evidence_urls":["https://..."]}]
  },
  "warnings": ["insufficient evidence for 2/9 claims"]
}
```

5.2 Tunable Parameters

- Retriever: FAISS/HNSW (e.g., HNSW32), k=20; minimum BM25 overlap=3.
- NLI model: large transformer (max seq len 512), batched inference.
- Fusion: Logistic Regression (L2=1.0) or XGBoost (max_depth=4).
- Calibration: isotonic on FEVER(dev).
- Thresholds: optimize Youden’s J or fix Precision@High ≥ 0.9.

5.3 Observability

- Log per-claim entailment, source priors, and final feature vectors.

- Weekly calibration drift reports (ECE, Brier).
- Alerts on entailment-rate drops or contradiction spikes.

6. Maintenance & Model Governance

- Update source priors weekly via offline KBT-style jobs; cache online lookups.
- Review hard rules quarterly (map to E-E-A-T/NewsGuard checklists).
- Dataset hygiene: versioned snapshots of FEVER/LIAR/MultiFC; guard against domain leakage; rotate holdouts.
- Fairness & abuse resistance: if crowd ratings are used, adopt bridging-based aggregation and monitor factional divergence.
- Security/Privacy: PII scrubbing on logs; hash URLs; respect robots.txt.

7. Roadmap for Algorithmic Improvements

- Hybrid dense-sparse retrieval; hard-negative training for stronger evidence recall.
- Blend external ratings (e.g., NewsGuard) as weak labels; boost scholarly sources for scientific topics.
- Temporal robustness: penalize stale citations; add publication-to-event latency features.
- Conformal prediction for uncertainty bands.
- Domain-specific adapters for health/finance entailment.
- Human-in-the-loop: rater checklist when scores fall into an uncertainty band (e.g., 0.45–0.65).

8. Reproducibility Checklist

- Fixed random seeds; reproducible environment file; model/index checksums.
- Data cards for FEVER, LIAR, MultiFC with citation and split details.
- Versioned configs for calibration and thresholds.
- Scripts: prepare_data/, train_fusion/, eval/, calibrate/, ablations/.
- Notebook template for reliability diagrams and ablation tables.

9. References (Selected)

- Dong et al. (2015). Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. PVLDB.
- Thorne et al. (2018). FEVER: Fact Extraction and VERification. NAACL.
- Wang (2017). LIAR: A New Benchmark Dataset for Fake News Detection. ACL.

- Augenstein et al. (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking. EMNLP.
- Shu et al. (2017). Fake News Detection on Social Media: A Data Mining Perspective.
- ACM Computing Surveys (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Challenges.
- Castillo et al. (2011). Information Credibility on Twitter. WWW.
- Google Search Quality Rater Guidelines (E-E-A-T).
- NewsGuard Rating Process & Criteria.
- Community Notes: Bridging-Based Ranking (overview and public docs).